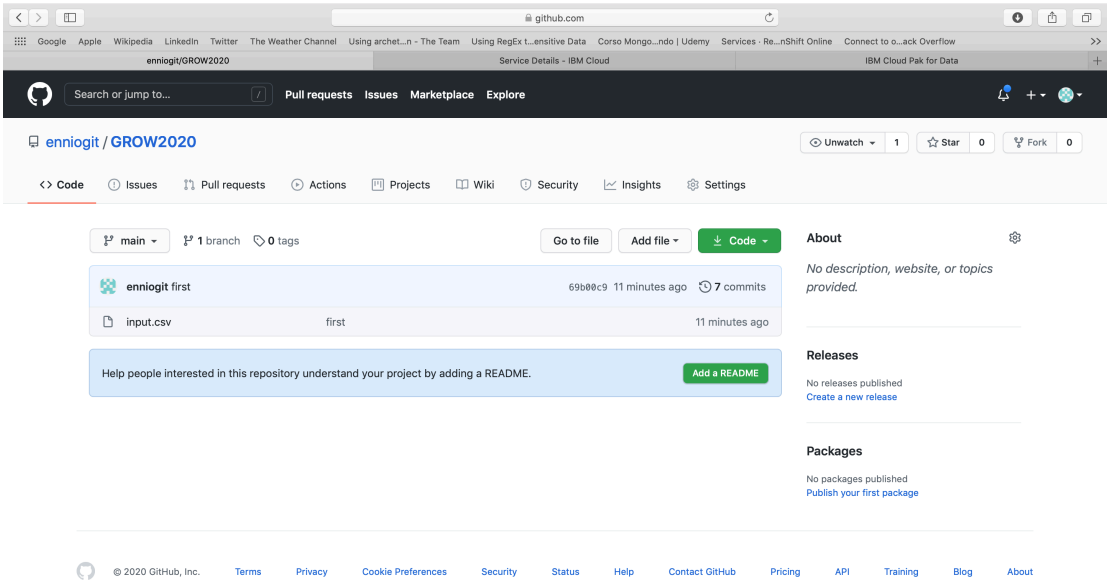
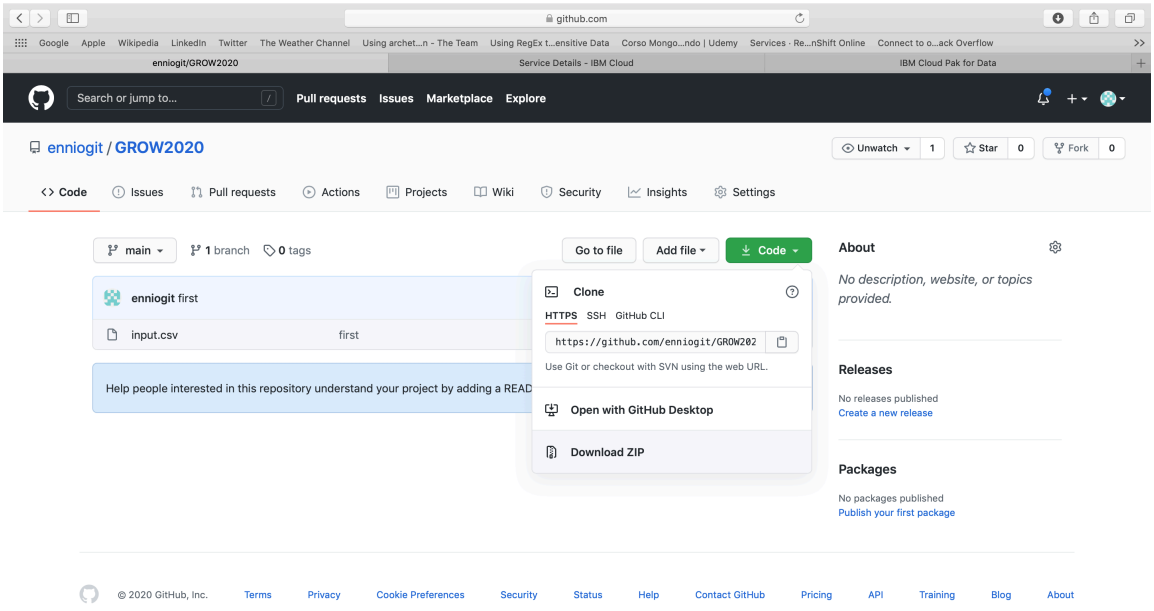
Go to
https://github.com/enniogit/GROW2020

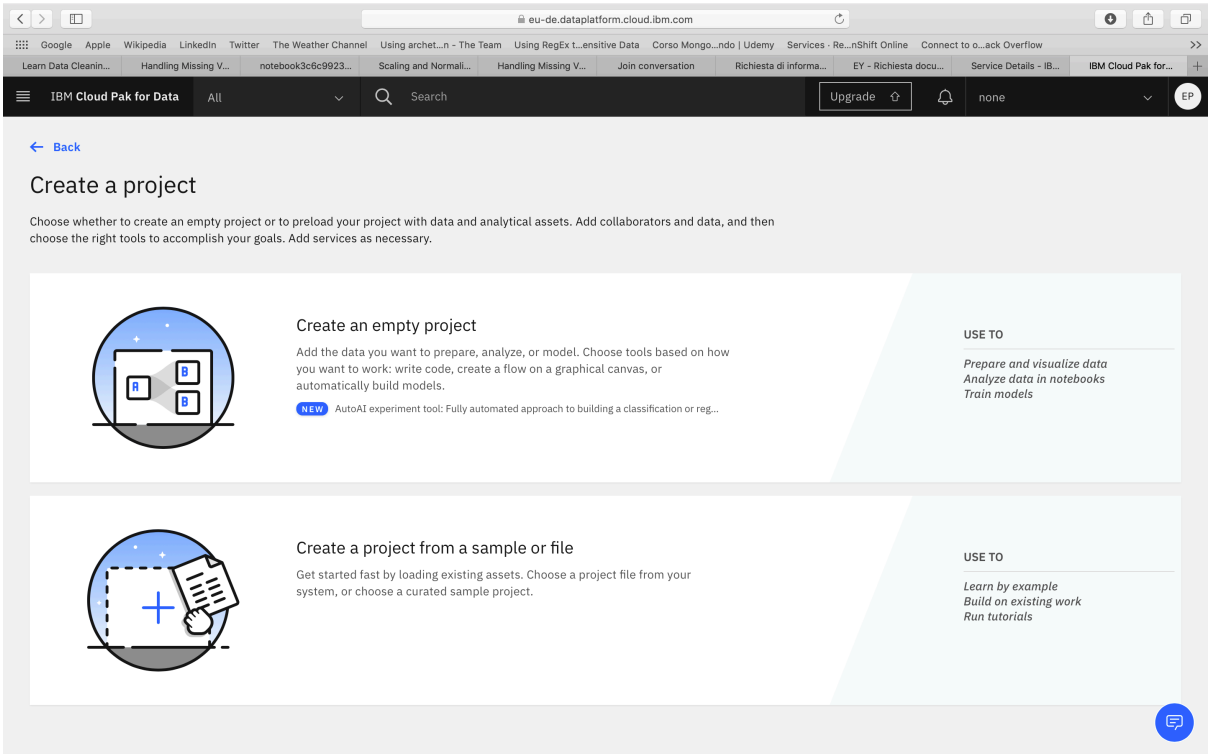


Select Code and then download zip

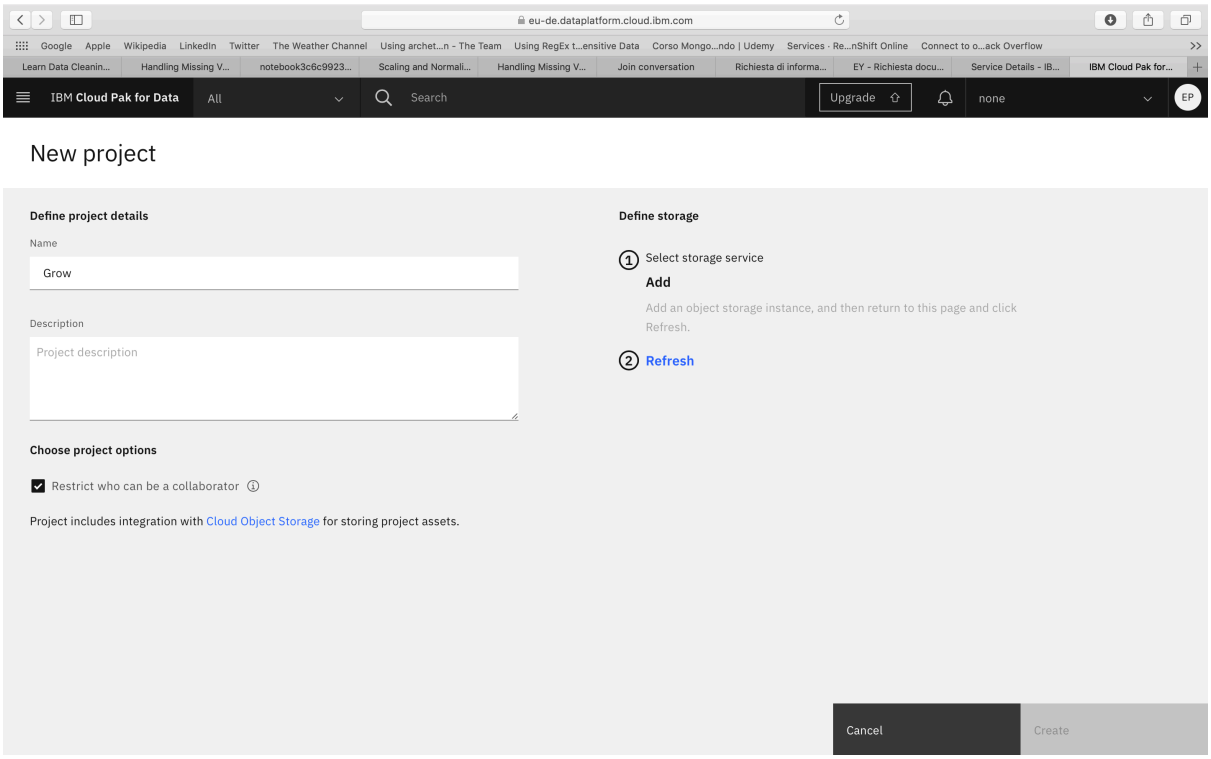


After, unzip the file and have a look at the folder.
There will be some files included this document. We will see them later.

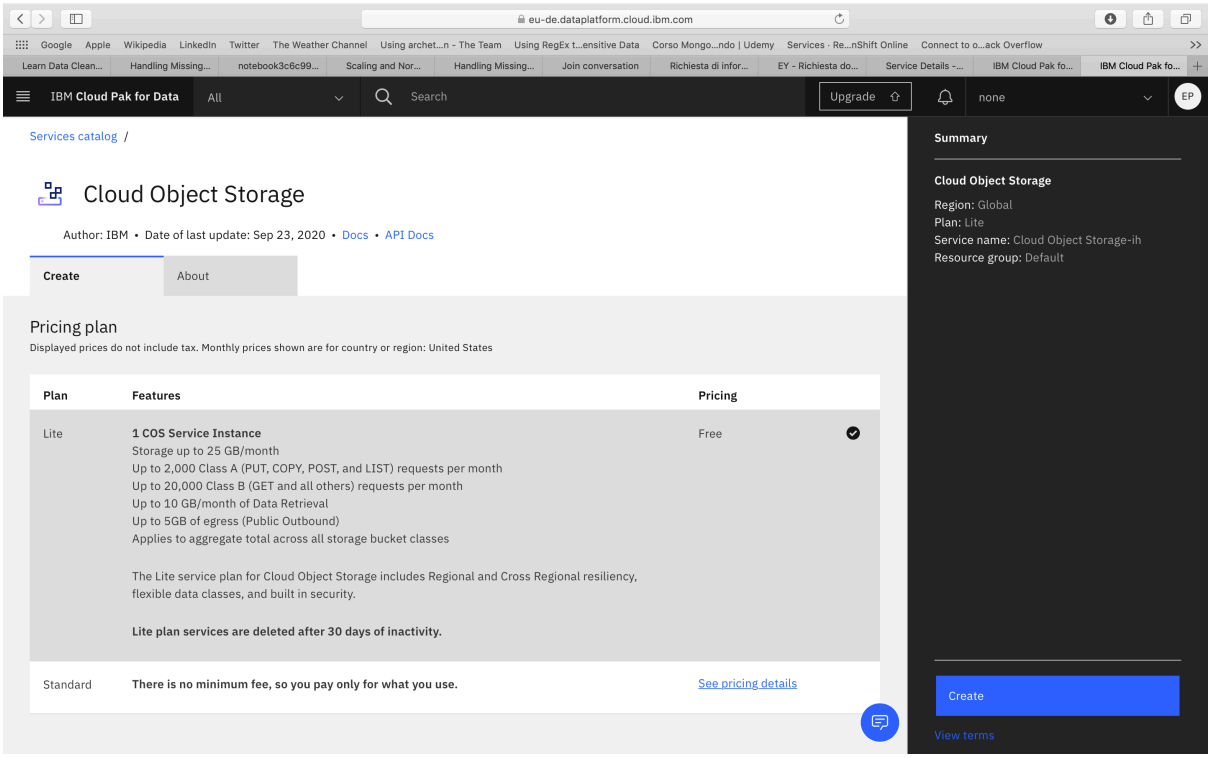
Go to https://dataplatform.cloud.ibm.com/
After login to be done by using the credential of registration you will get this screen



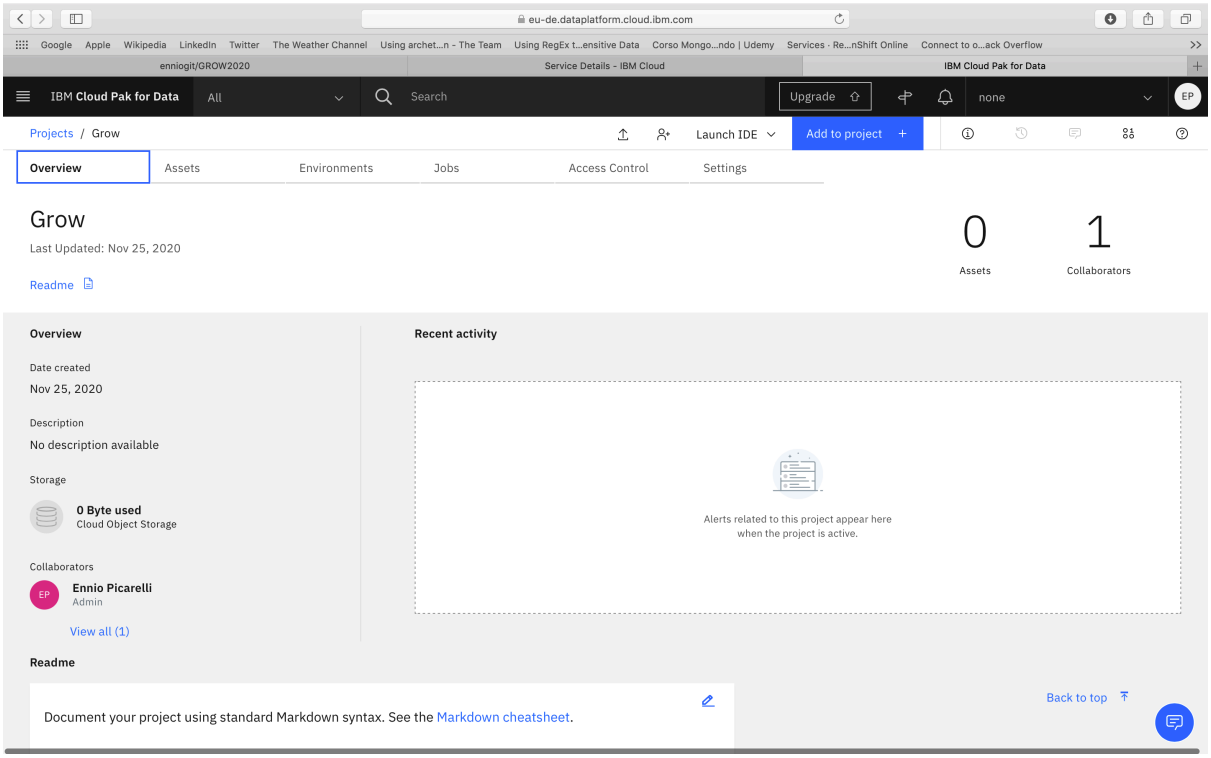
Select Create an empty project



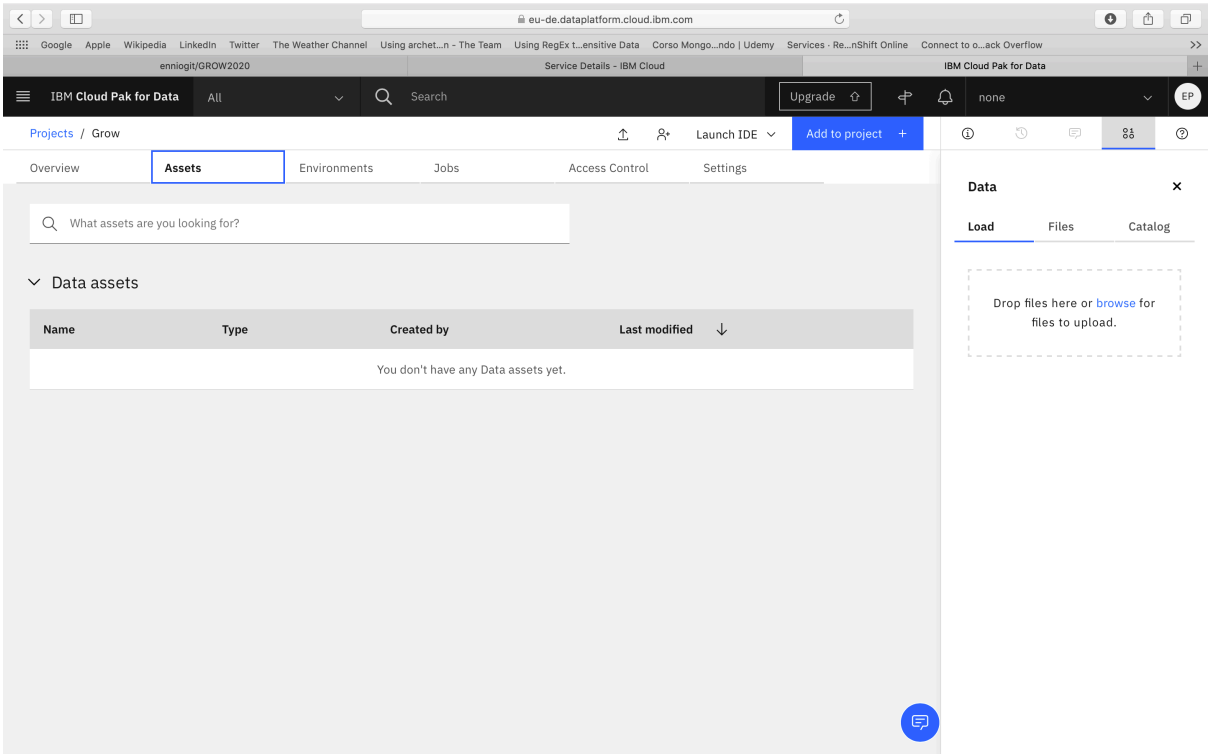
And then insert the name of the project and select the Add link in the top right of the screen.
You'll get the following



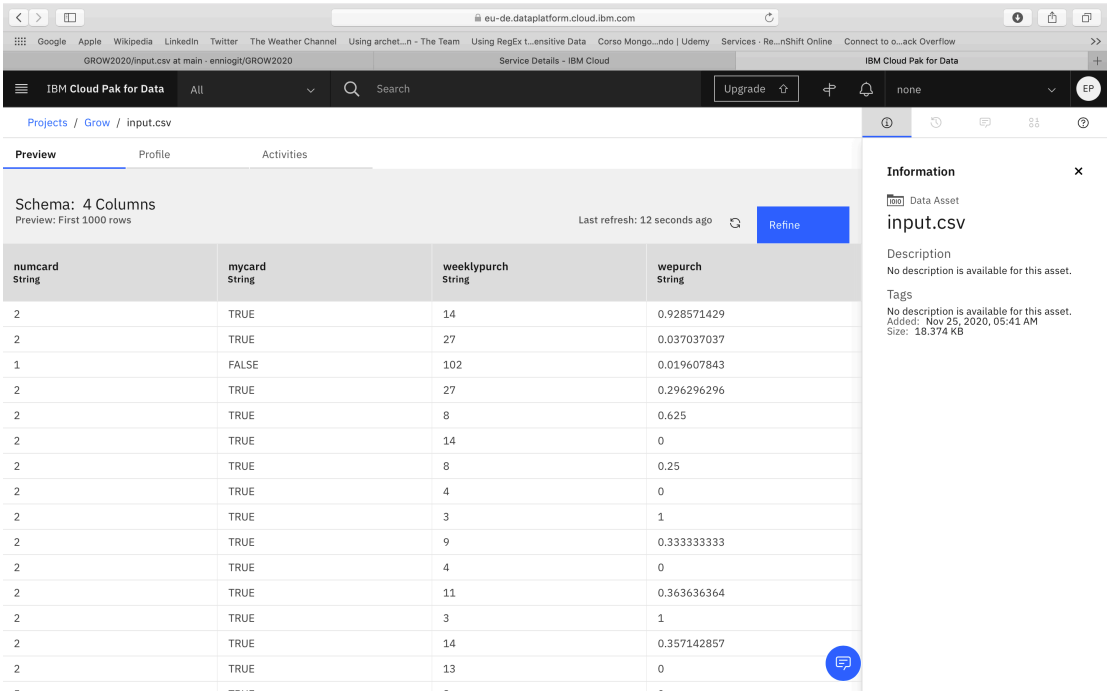
Push the Create button.
After project creation You'll get the following screen.



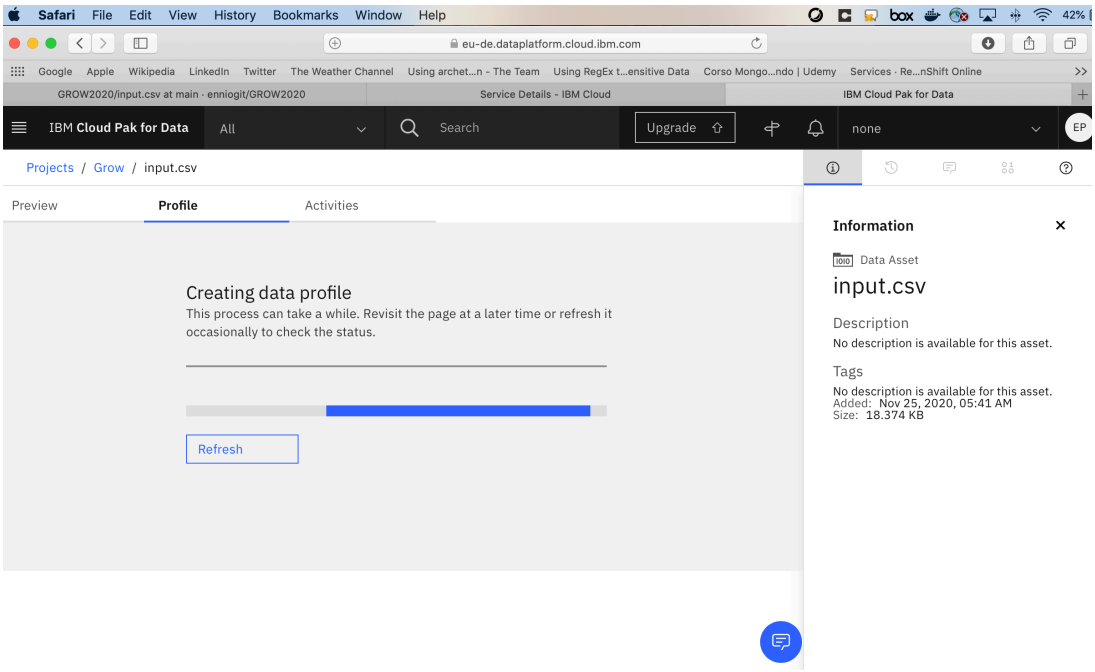
Select Asset



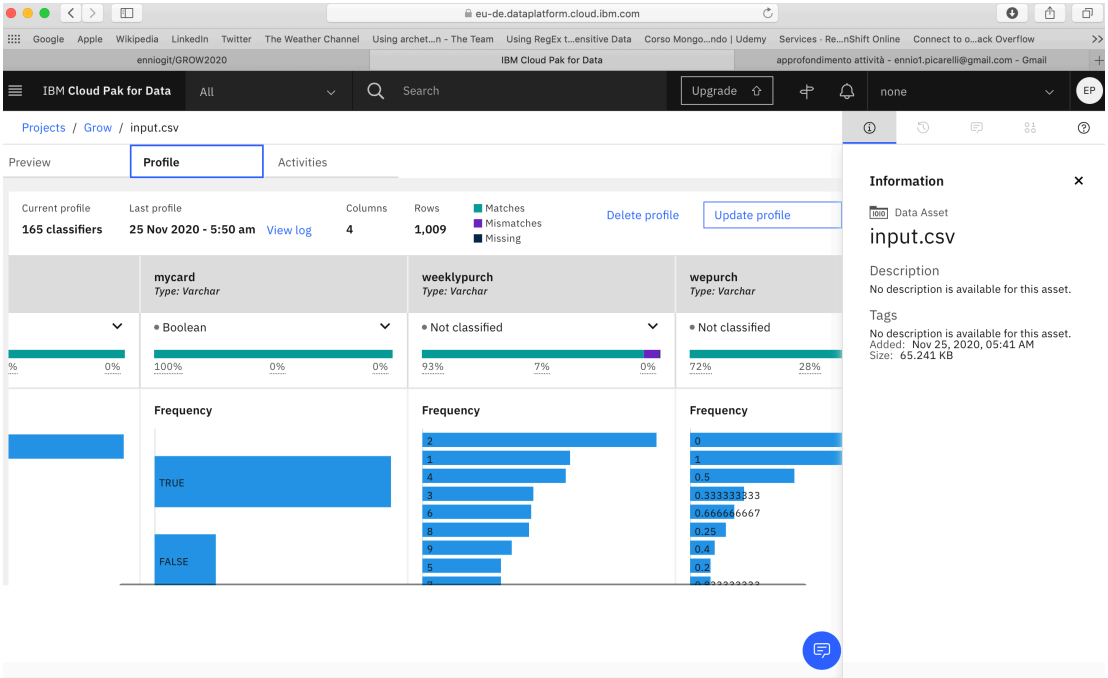
Then via drag & drop upload the input.csv file and have a look at it.



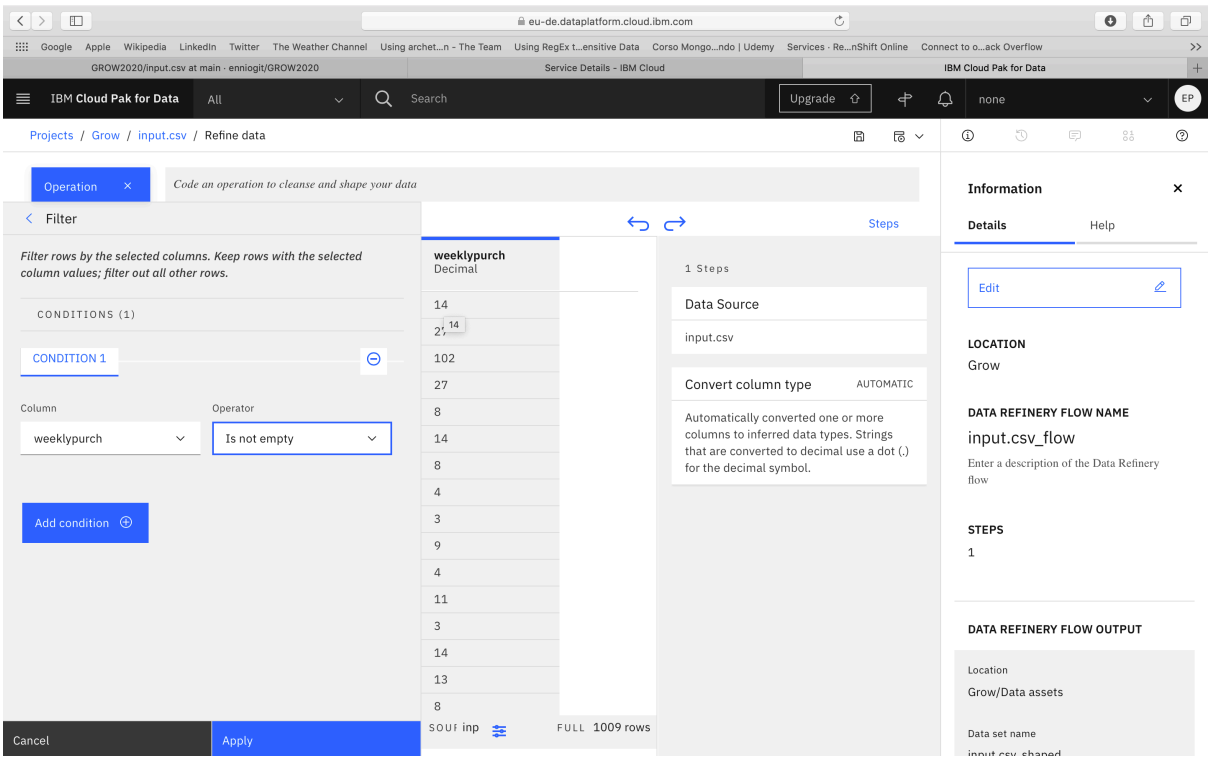
In order to quickly analyze the content of the dataset, you can perform the Profile task (it will get a while).



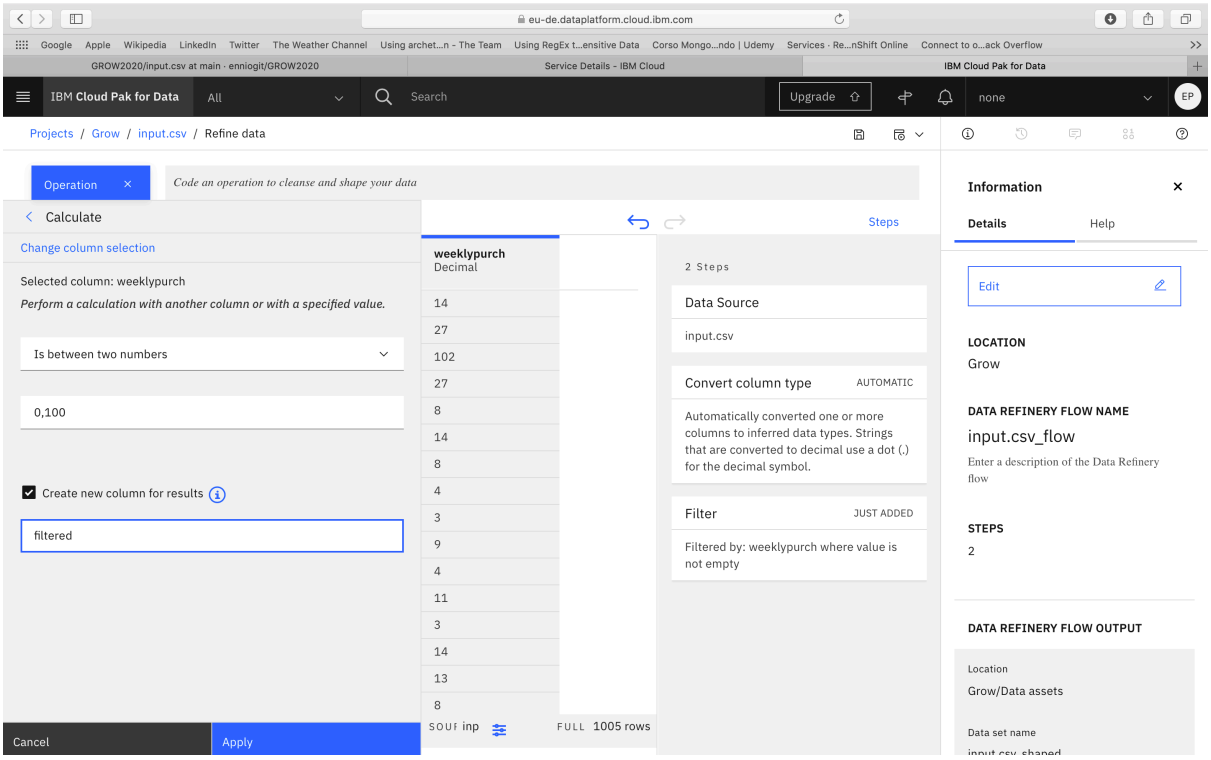
If you have a look at the profile results you can see that the weeklypurch has got some raw with null values and some other outliers (values below 0 or very high)



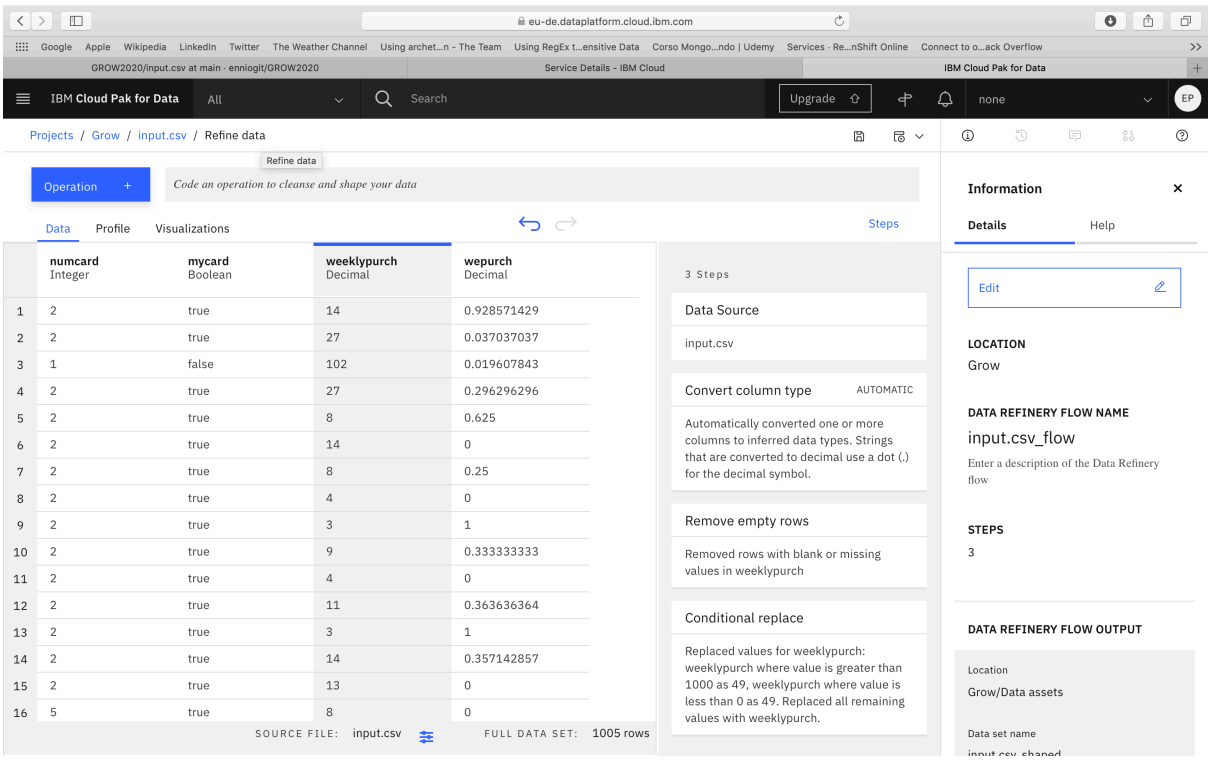
You can clean the dataset by using the refine function.
If you select Operation you can filter the raw corresponding to the missed values.



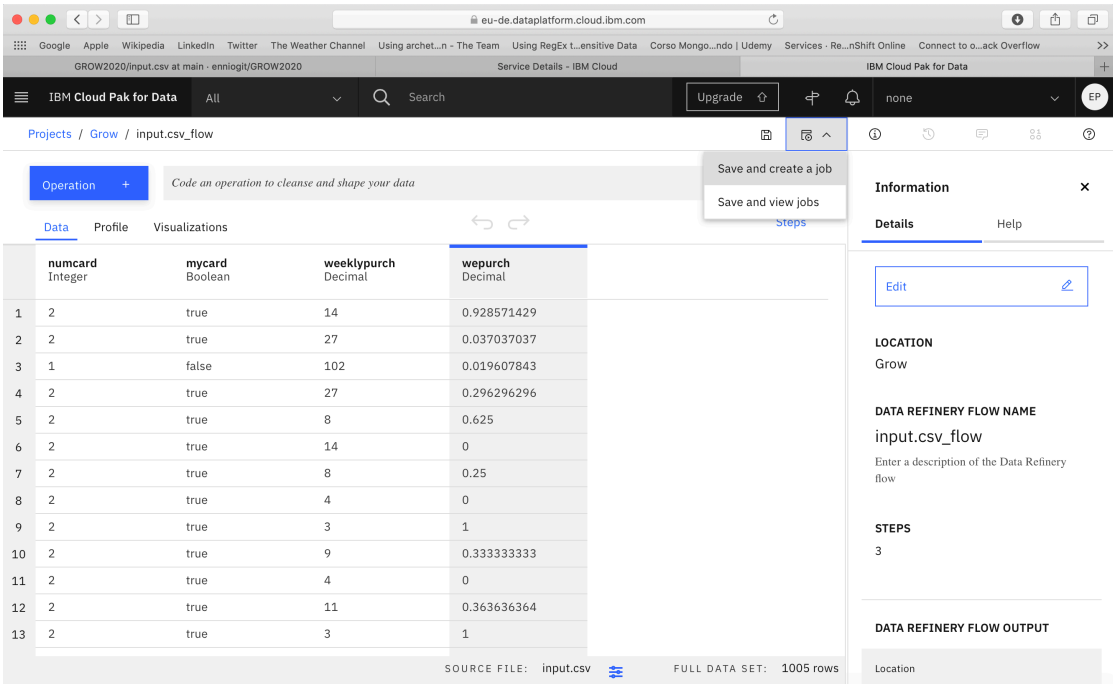
And you can replace withe the 49 value (corresponding to the column mean) the negative or the values higher than 200.

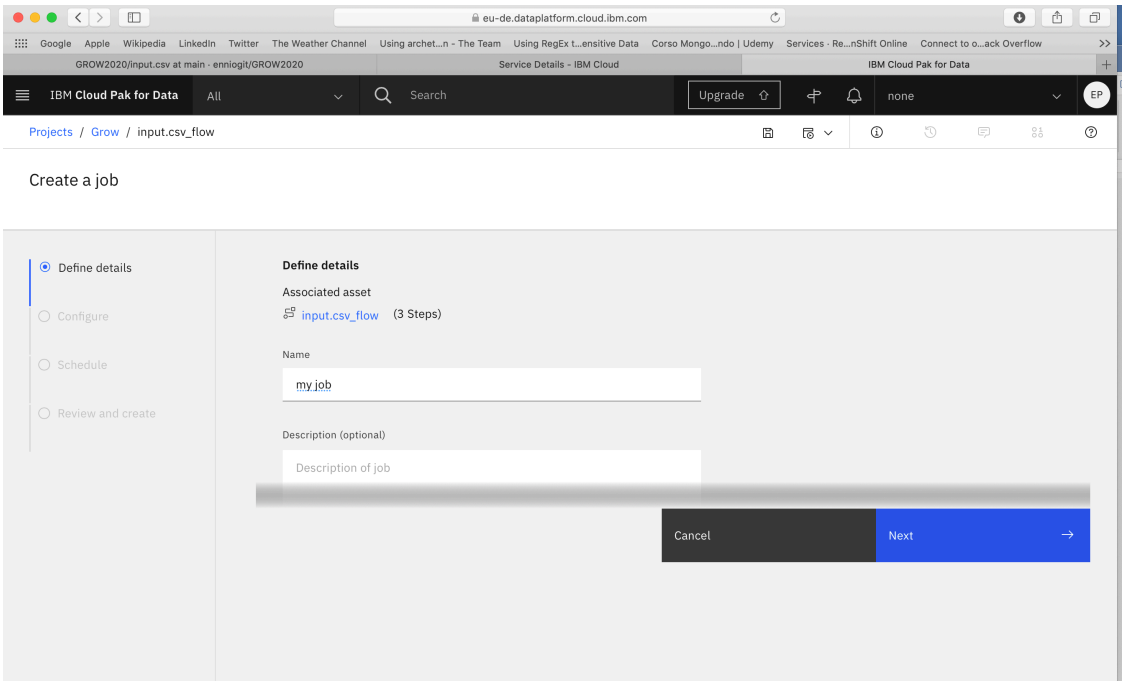


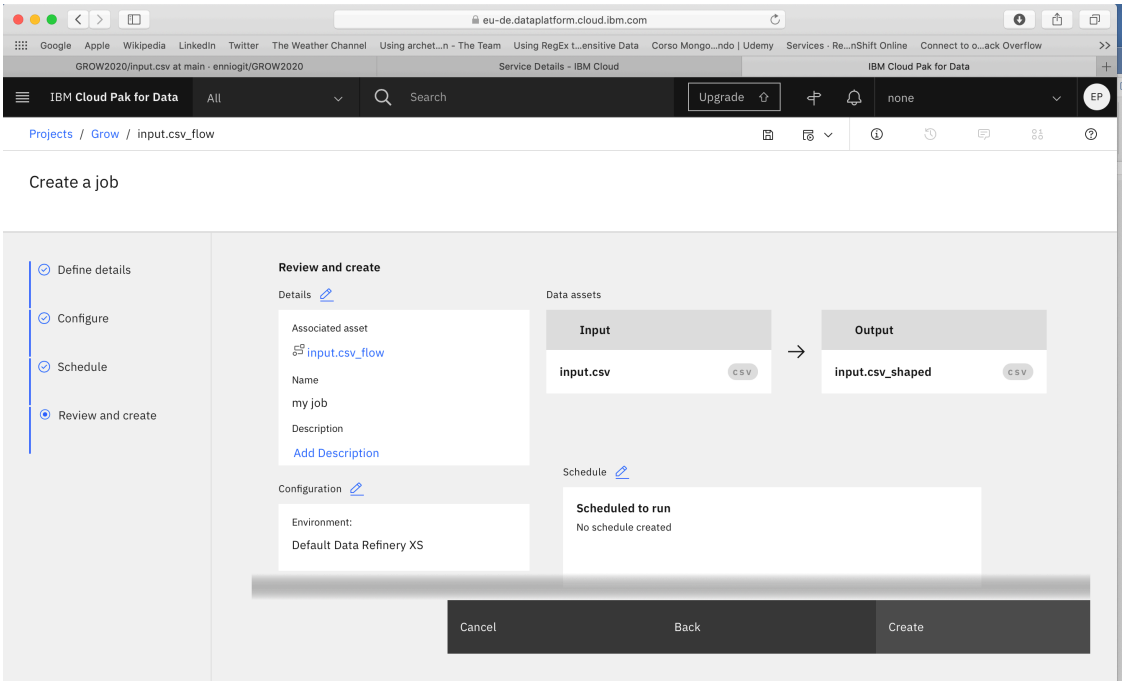
After this you have created a data refinery flow

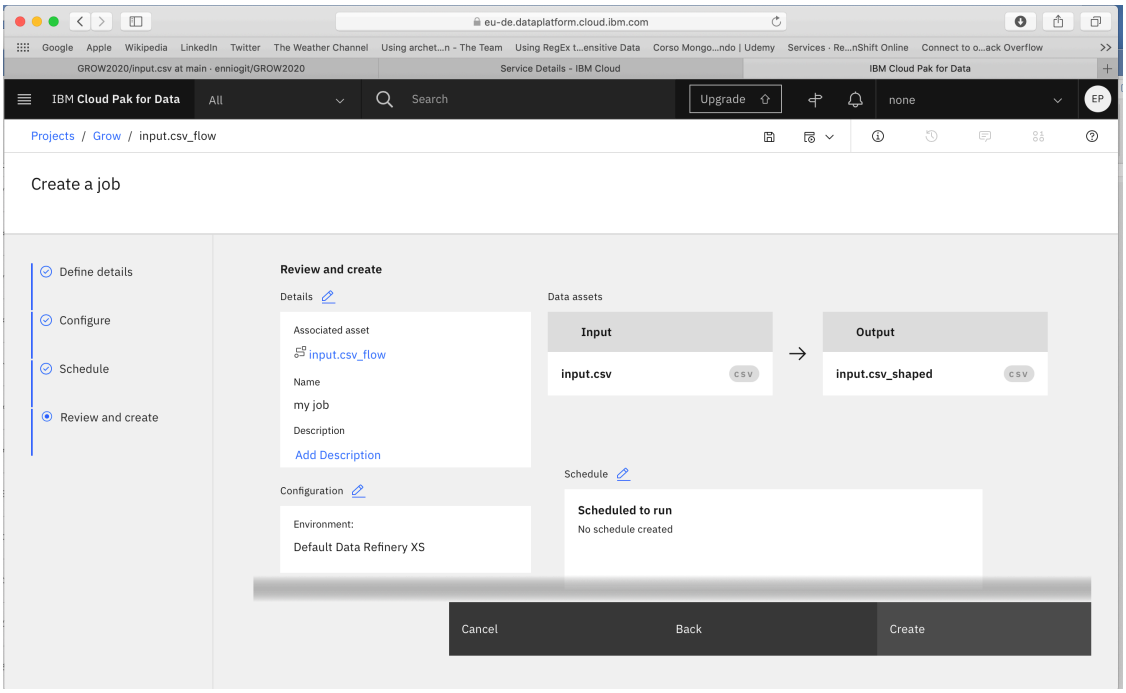


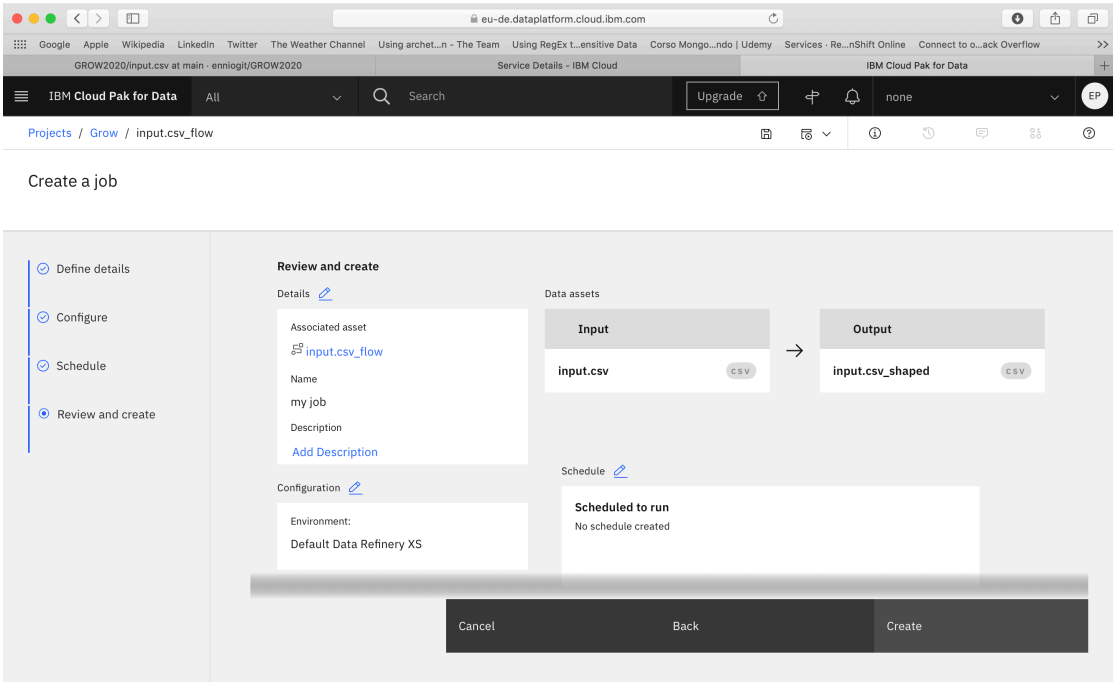
And you can save and create a job



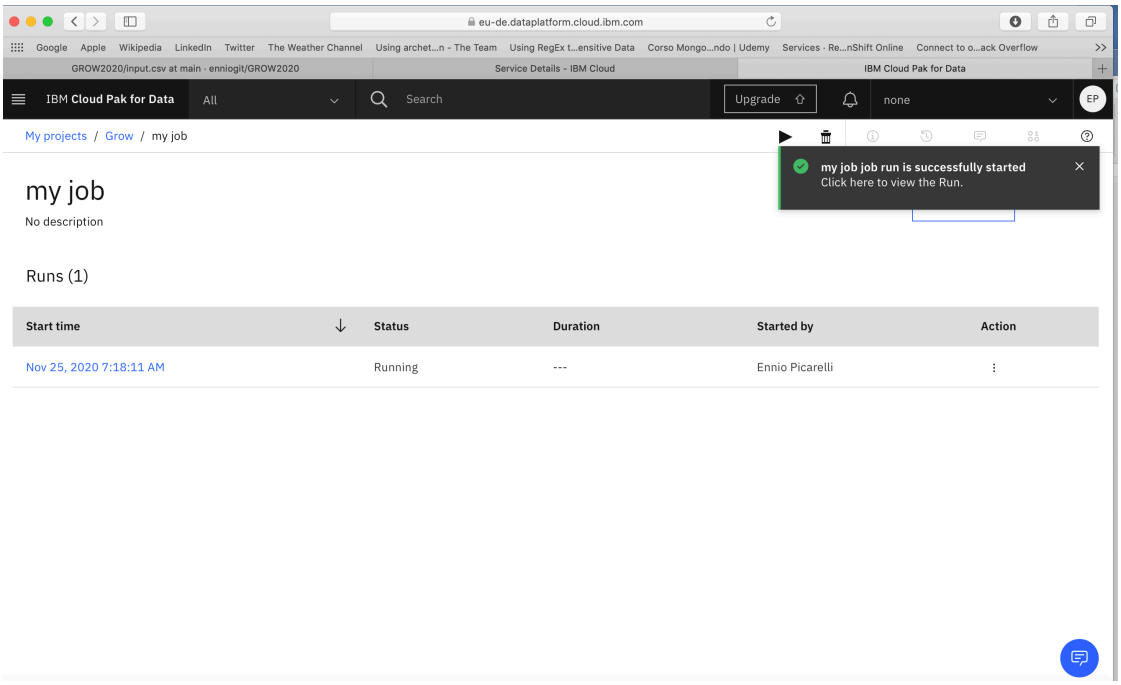

eu-de.dataplatform.cloud.ibm.com
IBM Cloud Pak for Data
All
Search
Upgrade
none
EP
Projects / Grow / input.csv_flow
Create a job
Define details
Configure
Schedule
Review and create
Define details
Associated asset
input.csv_flow (3 Steps)
Name
my job
Description (optional)
Description of job
Cancel
Next


eu-de.dataplatform.cloud.ibm.com
IBM Cloud Pak for Data
All
Search
Upgrade
none
EP
Projects / Grow / input.csv_flow
Create a job
Define details
Configure
Schedule
Review and create
Review and create
Details
Associated asset
input.csv_flow
Name
my job
Description
Add Description
Configuration
Environment:
Default Data Refinery XS
Data assets
Input
input.csv
CSV
Output
input.csv_shaped
CSV
Schedule
Scheduled to run
No schedule created
Cancel
Back
Create

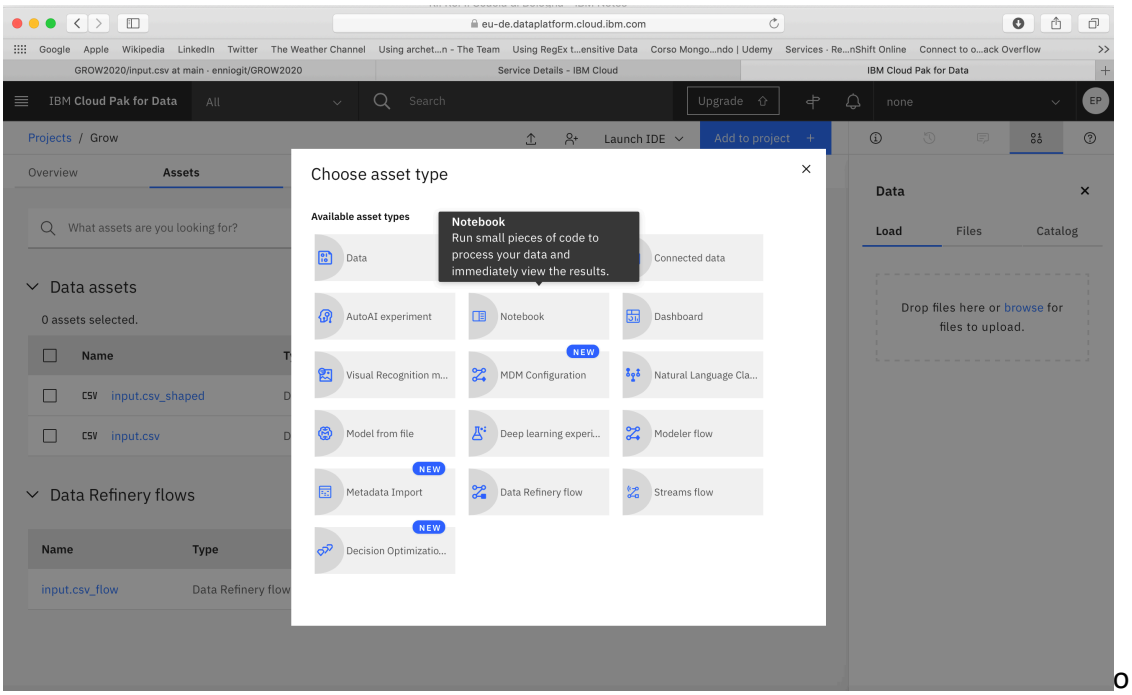




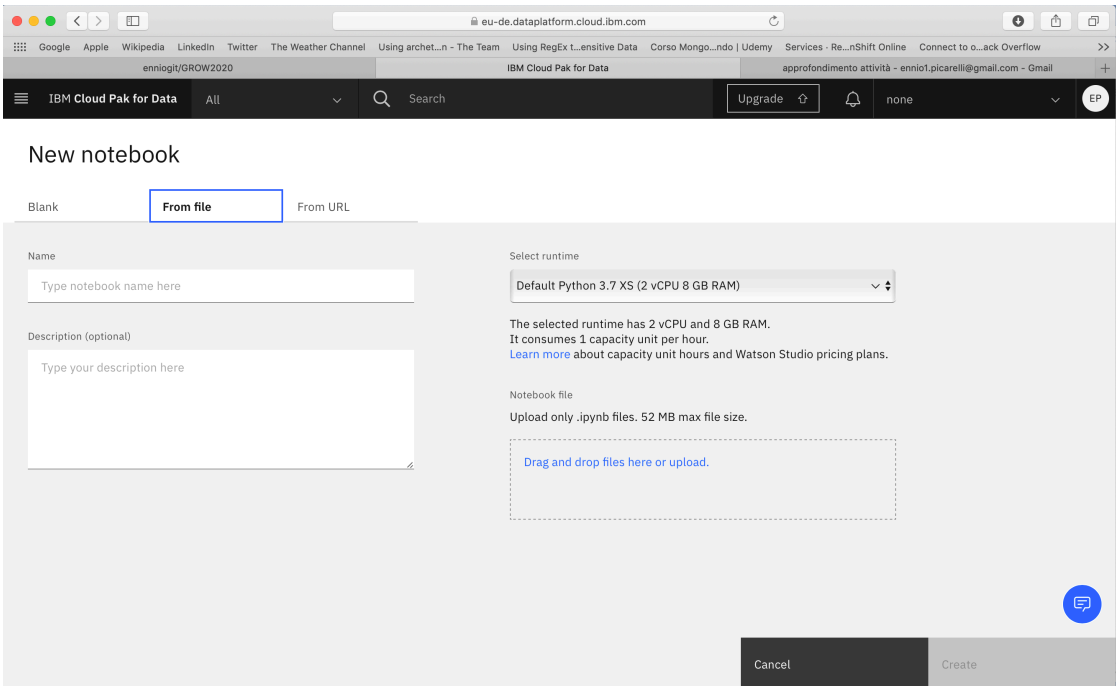
This job can then be executed immediately or scheduled



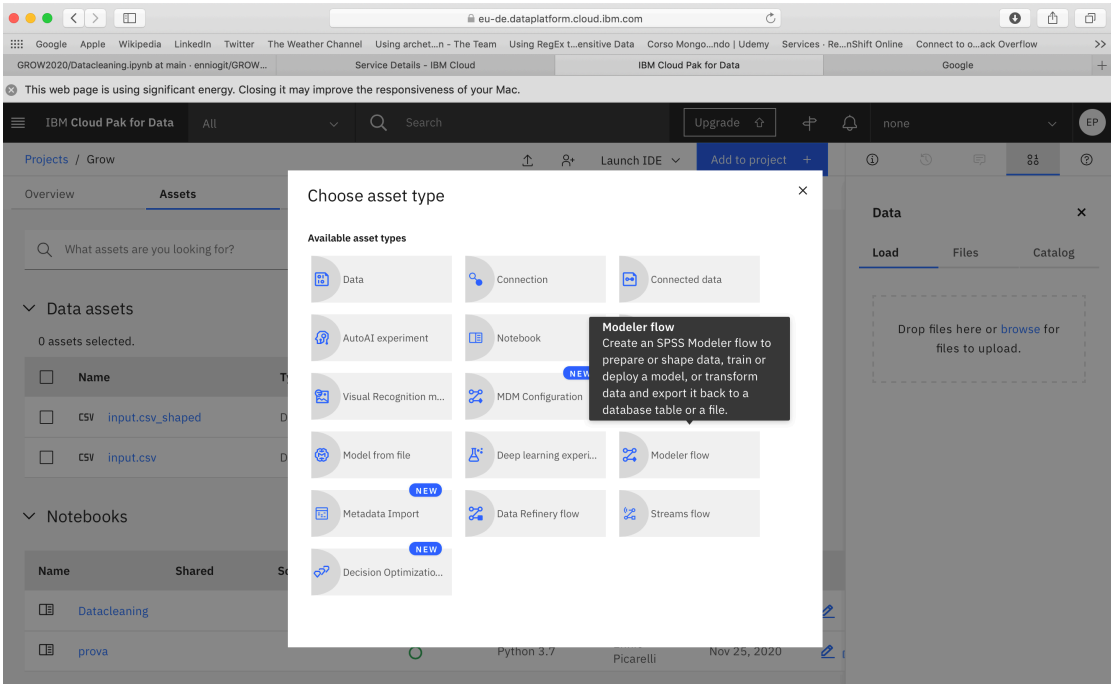
Now select add to project and from the popup Notebook

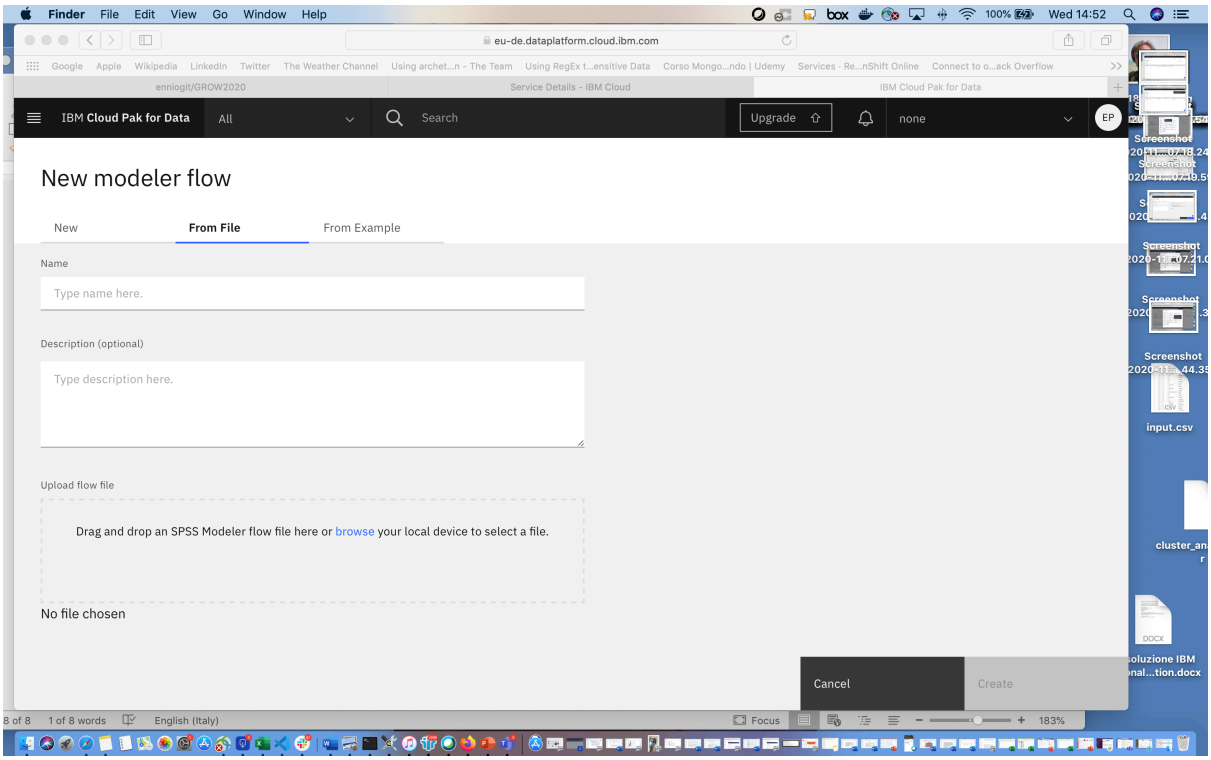


This is the environment that allows you to execute Jupyter notebooks with Python, R and Spark capabilities.
You can create a new Notebook and select the From file option then you can drag&drop the Datacleaning.ipynb file.



After the Notebook you can have a look as well at the modeler flow





Drag & drop Datacleaning.str and then select create.
Repeat the same operation with cluster_analysis